An Efficient Load Balancing Algorithm for virtualized Cloud Data Centers

Ali Naser Abdulhussein Abdulhussein, Jugal Harshvadan Joshi,
Atwine Mugume Twinamatsiko, Arash Habibi Lashkari, Mohammad Sadeghi

Postgraduate Centre of Study (PGC),
Limkokwing University of Creative Technology, Cyberjaya, Malaysia

Abstract: Cloud computing has become a new computing paradigm as it can provide scalable IT infrastructure, QoS-assured services and customizable computing environment. Although there are many research activities or business solutions for Cloud computing, most of them are focused on single-provider Cloud. As a key service delivery platform in the field of service computing. Cloud Computing provides environments to enable resource sharing in terms of scalable infrastructures, middleware and application development platforms, and value-added business applications .This study examined the latest technology in the field Cloud Computing. The main study focused on load balancing for virtual machines inside single cloud data center. There different algorithms for balancing, one of them called Throttled load balancing which treats the virtual machines based on two values that can send to the intended virtual machine or send it to the remote ones .A proposed modification has been proposed to solve some of the key features in this algorithm like Process migration, Fault tolerant and Overload Rejection. The idea is to send even when all the virtual machines heavily loaded by determining the most respectable hardware specifications of the virtual machines.

Keywords: Cloud Computing, cloud Virtualization, load balancing, Cloud Data Center

Introduction

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services. With a connection over the internet, a consumer is able to access various resources, be it premium or free in order to perform certain functionality and all these constitute a cloud; The services themselves have long been referred to as Software as a Service (SaaS). Some vendors use terms such as IaaS (Infrastructure as a Service) and PaaS (Platform as a Service) to describe their products. This is to say with cloud computing, a cloud is formed over the amalgamation of various services be it physical or virtual over a network to perform certain services (Parhizkar B. et.al, 2013).

By deploying IT infrastructure and services over the network, an organization can purchase these resources on an as needed basis and avoid the capital costs of software and hardware. With cloud computing, IT capacity can be adjusted quickly and easily to accommodate changes in demand. While remotely hosted, managed services have long been a part of the IT landscape, a heightened interest in cloud computing is being fueled by ubiquitous networks, maturing standards, the rise of hardware and software virtualization, and the push to make IT costs variable and transparent.

Great interest in cloud computing has been manifested from both academia and private research centers and numerous projects from industry and academia have been proposed. In commercial contexts among the others, we highlight: amazon elastic compute cloud, IBM's blue cloud, etc .There are several scientific activities driving toward open cloud-computing middleware and infrastructures such as reservoir and eucalyptus, etc (Parhizkar B. et.al, 2013).

Clouds aim to power the next generation data centers as the enabling platform for dynamic and flexible application provisioning. This is facilitated by exposing data center's capabilities as a network of virtual

services (e.g. Hardware, database, user-interface, and application logic) so that users are able to access and deploy applications from anywhere in the Internet driven by the demand and QoS (Quality of Service) requirements. Similarly, IT companies with innovative ideas for new application services are no longer required to make large capital outlays in the hardware and software infrastructures. By using clouds as the application hosting platform, IT companies are freed from the trivial task of setting up basic hardware and software infrastructures. Thus they can focus more on innovation and creation of business values for their application services.

Related works

According to Abhay Bhadani and Sanjay Chaudhary (Bhadani A. et.al, 2010), they propose a Central Load Balancing Policy for Virtual Machines (CLBVM) to balance the load evenly in a distributed virtual machine/cloud computing environment. This effort tries to compare the performance of web servers based on their CLBVM policy and independent virtual machine(VM) running on a single physical server using Xen Virtualizaion. The paper discusses the value and feasibility of using this kind of policy for overall performance improvement.

The proposed CLBVM policy is at an inception stage and requires work for I/O as well as memory availability on the target server. The CLBVM policy has the potential to improve the performance of the overall N Servers though it does not consider fault tolerant systems. They tried to make the system completely distributed such that, if the performance of the VM gets affected by another VM, it can move itself to lightly loaded server on the go (Bhadani A. et.al, 2010).

Glauco Estácio Gonçalves et. al (Gonçalves, G. E. et.al 2013), proposed algorithms for allocation of computing and network resources in a Distributed cloud (D-Cloud) with the objectives of balancing the load in the virtualized infrastructure and of considering constraints, such as processing power, memory, storage, and network delay. The evaluation of the algorithm shows that it is indeed adequate for link allocation across different physical networks. It considers that links are unconstrained in terms of capacity. They argue that this situation is well-suited to a pay-as-you-go business plan, very common in Cloud Computing and it allows a better usage of the resources than the common idea of link capacity reservation. The proposed algorithms were tested through simulations, focusing on the improvements brought by the minimax path strategy. The experiments showed that the minimax path strategy can offer better load balancing, in terms of maximum link stress than heuristics from the literature as the rate of requests increases (Gonçalves, G. E. et.al, 2013).

Zehua Zhang and Xuejie Zhang (Zhang Z. et.al , 2010), proposed a load balancing mechanism based on ant colony and complex network theory in open cloud computing federation in this paper, it improves many aspects of the related Ant Colony algorithms which proposed to realize load balancing in distributed system, Furthermore, this mechanism take the characteristic of Complex Network into consideration. Finally, the performance of this mechanism is qualitatively analyzed, and a prototype is developed to enable the quantitative analysis, simulation results manifest the analysis.

This mechanism improves many aspects of the related Ant Colony algorithms which proposed to realize load balancing in distributed system, and the characteristic (small-world and scale-free) of Complex Network have been taken into consideration (Zhang Z. et.al , 2010).

Srinivas Sethi et. al (Sethi S. et.al , 2012) , they introduced the novel load balancing algorithm using fuzzy logic in cloud computing, in which load balancing is a core and challenging issue in Cloud Computing. The processor speed and assigned load of Virtual Machine (VM) are used to balance the load in cloud computing through fuzzy logic. It is based on Round Robin (RR) load balancing technique to obtain measurable improvements in resource utilization and availability of cloud-computing environment.

The network structure or topology also required to take into consideration, when creating the logical rules for the load balancer. Two parameters named as the processor speed and assigned load of Virtual Machine (VM) of the system are jointly used to evaluate the balanced load on data centers of cloud computing environment through fuzzy logic (Sethi S. et.al, 2012).

Pengfei Sun et. al, They presented a new security load balancing architecture-Load Balancing based on Multilateral Security (LBMS) which can migrate tenants' VMs automatically to the ideal security physical machine when reach peak-load by index and negotiation. They have implemented their prototype based on CloudSim, a Cloud computing simulation. Their architecture makes an effort to avoid potential attacks when VMs migrate to physical machine due to load balancing (Sun P. et.al, 2011).

Shu-Ching Wang et.al (Wang S. et.al, 2010), they introduced a two-phase scheduling algorithm under a three-level cloud computing network is advanced. The proposed scheduling algorithm combines OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms that can utilize more better executing efficiency and maintain the load balancing of system.

The goal of this study is to reach load balancing by OLB scheduling algorithm, which makes every node in working state. Besides, in their research, the LBMM scheduling algorithm is also utilized to make the minimum execution time on the node of each task and the minimum whole completion time is obtained. However, the load balancing of three-level cloud computing network is utilized, all calculating result could be integrated first by the second level node before sending back to the management. Thus, the goal of loading balance and better resources manipulation could be achieved (Wang S. et.al, 2010).

He-Sheng WU et.al (Wu H. et.al, 2013), discussed the new characteristics the load balancing should have in cloud computing. In cloud computing, load balancing manages virtual machine in the cloud instead of actual one. Therefore, load balancing system should be provided with the function of elastic management of back-end resource, i.e. to dynamically add or delete back-end server (existing in the form of virtual machine in the cloud) based on actual network load condition.

Since the virtual machine for load balancing management in cloud computing can be dynamically applied and released, an algorithm of prediction-based elastic load balancing resource management (TeraScaler ELB) is presented to overcome the drawbacks.

Experiments have shown that the required number of virtual machines change in compliance with the change of network load, thus TeraScaler ELB is able to dynamically adjust the processing capacity of back-end server cluster with the applied load. Besides it could make full use of the 'use on demand' feature of cloud computing, TeraScaler ELB leads to a better application of prediction based load balancing in cloud computing. It concludes that compared with the traditional elastic resource management algorithm, TeraScaler ELB is more reasonable for providing scalability and high availability (Wu H. et.al, 2013).

Jiann-Liang Chen et. al (Chen J. et.al , 2012), this paper presents a study to improve cloud computing systems performance based on Eucalyptus cloud platform. An optimal load balancing mechanism called EuQoS system for scheduling VMs is proposed. Extending EuQoS to accommodate real-time services, Hadoop platform is integrated into the EuQoS system. Log processing services are utilized to investigate the performance of system throughput. Experimental results indicate that the proposed EuQoS system can improve system throughput by 6.94% compared with the basic Eucalyptus platform with Hadoop mechanisms.

According to Xiaona Ren et. al (Ren X. et. al , 2011), Considering the unique features of long-connectivity applications, an algorithm is proposed, Exponential Smoothing forecast-Based on Weighted Least-Connection ESBWLC. ESBWLC optimizes the number of connections and static weights to actual load and service capability, and adds single exponential smoothing forecasting mechanism. Finally, experiments show that ESBWLC can improve the load of real servers effectively.

Analysis of Previous Load Balancing Algorithms

The throttled algorithm has been chosen to be modified based on the simulation conducted as it has better results compared with the other algorithms especially if it is working with response time algorithm and based on recommendations of the experts that we have found in the articles (Shiraz M. et. al , 2012). Throttled algorithm is also known as the threshold algorithm as it has two values (tUpper and tUnder) which specify the load on the virtual machines. If the load is greater than the value of tUpper then, it will send the process to the remote processor, otherwise if the load if less than the value of tUnder then it will process it locally and if the virtual machine is overloaded then it will update the other virtual machines of its state. The algorithm has a low inter-process communication as most of the load is processed locally which leads in performance as there is not much sending and receiving of jobs (Sanei Z. et. al, 2011).

There many metrics that govern the load balancing in a virtualized data centers, the threshold algorithm guarantees most of them except some which are as follows (Sharma S. et.al, 2008):

Overload Rejection: If Load Balancing is not promising additional overload rejection measures are needed. When the overload situation ends then first the overload rejection measures are stopped. After a short guard period Load Balancing is also closed down.

Fault Tolerant: This parameter gives that algorithm is able to bear twisted faults or not. It enables an algorithm to continue operating properly in the event of some failure. If the performance of algorithm decreases, the decrease is relational to the seriousness of the failure, even a small failure can cause total failure in load balancing.

Process Migration: Process migration parameter provides when does a system decide to migrate a process? It decides whether to create it locally or create it on a remote processing element. The algorithm is capable to decide that it should make changes of load distribution during execution of process or not.

Proposed New Algorithm

As a result we can conclude that when the virtual machines is heavily loaded and also the other virtual machines overloaded then it will not send any incoming jobs to the remote processor as it is overloaded and it will process it locally. our idea is when a virtual machines is overloaded, the algorithm can still send to the most respectable virtual machine in terms of physical hardware specifications, this will lighten the burden of one virtual machine to the other as it will make the strongest virtual machines handle the load while the others processing until the others will finish the jobs and back to the original state. This will help solving some the metrics that should be found in the load balancing algorithms like overload rejection (Abhijit A. et.al, 2012). This modification will make a queue of new incoming jobs for the strongest virtual machines respectively as the load on one overloaded processor can be much higher than on other overloaded processors, causing significant disturbance in load balancing, and increasing the execution time of an application.

Table 1: Analysis of Throttled algorithm (Abhijit A. et.al, 2012)

Parameters	Threshold Algorithm
Nature	Static
Overload Rejection	**Yes**
Reliability	Less
Adaptability	Less
Stability	Large
Predictability	More
Forecasting Agency	More

Cooperative	Yes
Fault Tolerant	**Yes**
Resource Utilization	Less
Process Migration	**Yes**
Preemptiveness	Non-preemptive
Response Time	Less
Waiting Time	More
Turnaround Time	Less
Execution System	Decentralized
Throughput	Low
Processor Thrashing	No

Simulation and Results

In this section, this paper will try to show some the simulations conducted to choose the throttled algorithm as the algorithm to modify. First simulation was conducted to test throttled algorithm with the closest data center as the service broker whereby it chooses the closest data center to the user request. The second simulation was conducted for the throttled algorithm but in regard to the best response time of the data centers the request was made to. In the third round the simulation was tested for throttled in regard to reconfiguring dynamically of the data center located around the world whereby the requests can be directed flexibly according to the best data center that can serve the request.

Throttled With Closest Data Center

The first round was tested using the Throttled algorithm for the load balancing strategy with the Closet Data Center as the Data Center broker policy. The simulations were conducted for 100 virtual machine and a total of 1000 requests per user per hour, with an average of 10000 users in a peak hours and 100 in off-peak hours with 6 user bases located in different locations around the world and 6 Data Centers also located in different locations around the world as shown in figure 1.

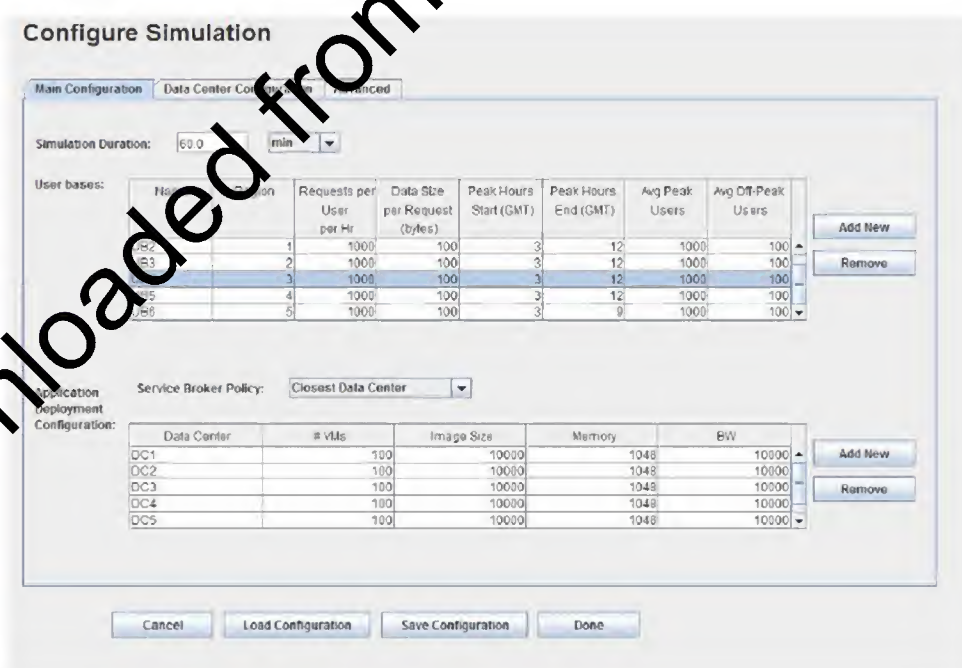

Figure 1: User base specification and service broker policy for the data center

In figure (2, 3), explanation on how to do the configuration for the simulation that is going to be tested on, Whereby the selection of the user requests that can be made per hour and whether it's a peak hour or normal hours to simulate a real world events as well as the starting and ending times for these requests. The data center broker that will govern the behavior of the data center have been identified as well ,for example in this simulation ,the closest data center which means the closest one for the requests to be entertained around the world. Hardware specifications of the data centers like the memory , CPU speed , the number of cores inside each CPU ,the operating system, the number of virtual machines inside each data centers and the cost that for each virtual machine to use the memory and CPU cores have been defined as well .

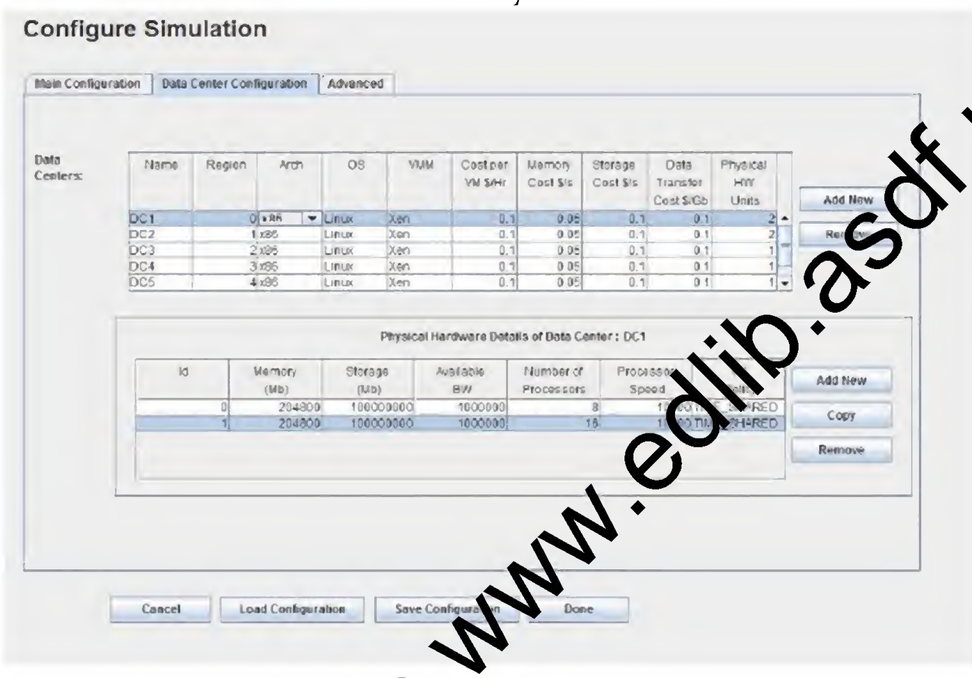

Figure 2: Data Center specifications

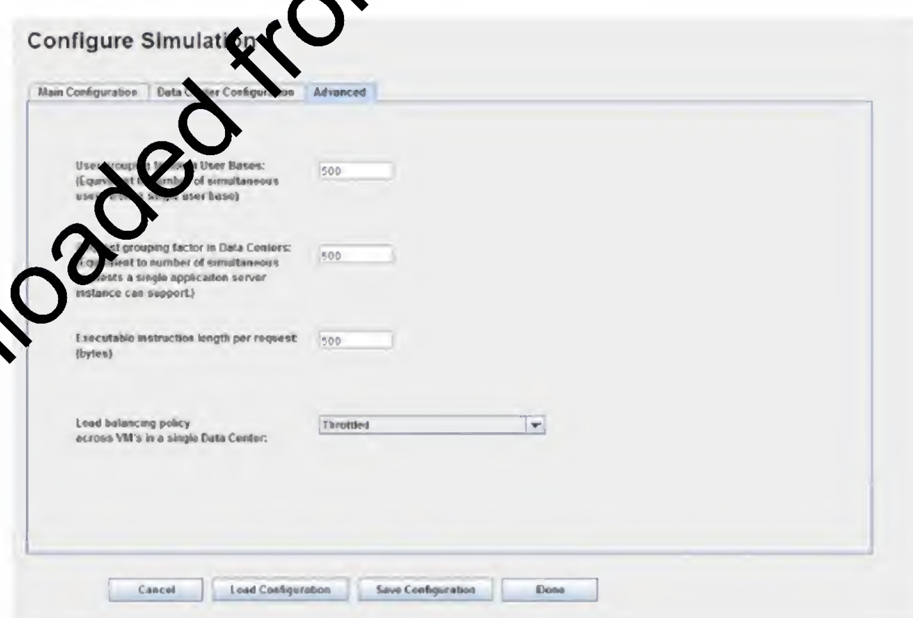

Figure 3: User requests and the load balancing algorithm

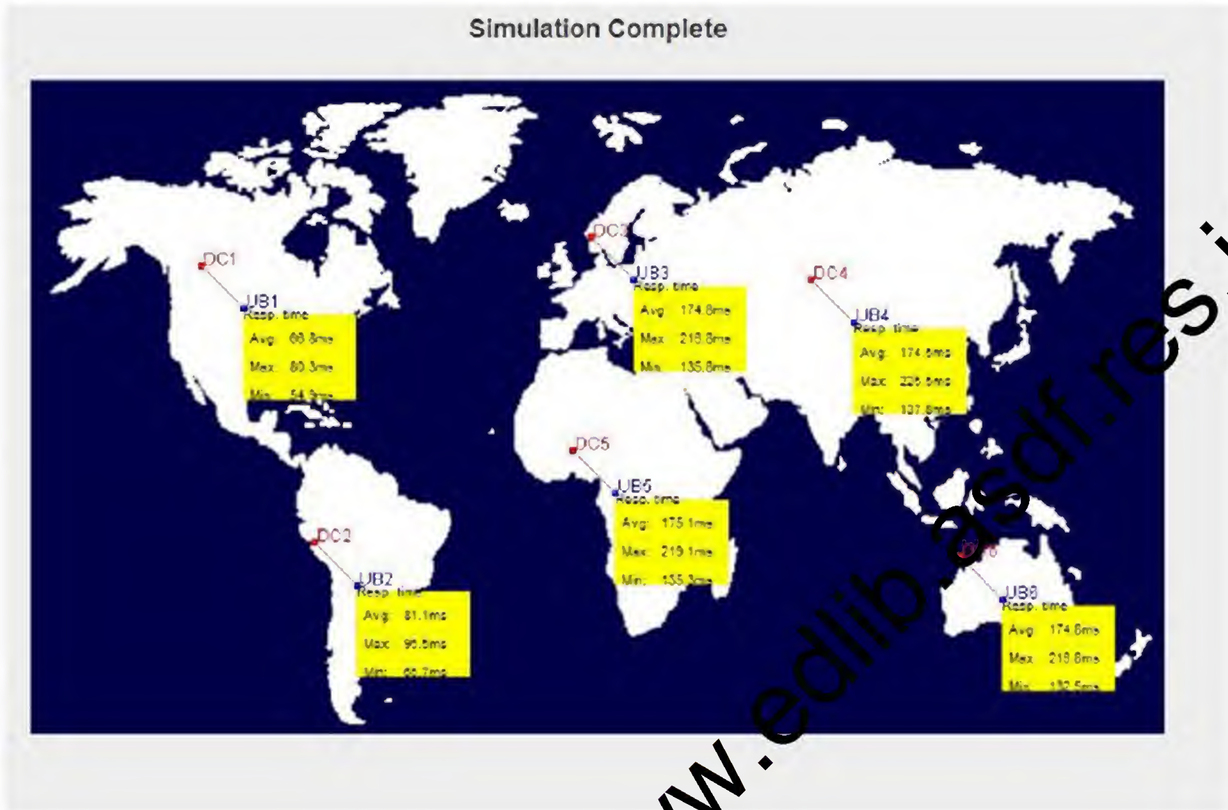

Figure 4: Data center locations and user bases requests

In figure 4, it shows the geographical location of each data center with their average, minimum and maximum execution time that can be taken in consideration when simulating with respect to each user request and whether this request is made from single user or multiple.
The results for the simulation conducted based on the specification provided in the above figures that shows the overall response time for the data centers and the cost for the virtual machines to serve the requests if the broker is set to closest data center.

Results

Avg (ms)	Min (ms)		Max (ms)
Overall response time	141.35	54.87	225.52
Data Center processing time	91.72	12.65	167.76

Cost

Total Virtual Machine Cost ($)	Total Data Transfer Cost ($)	Grand Total ($)
60.00	5.76	65.76

Throttled with Optimize Response Time

The second round was tested using the Throttled algorithm for the load balancing strategy with the Closet Data Center as the Data Center broker policy. I have conducted a simulation for 100 virtual machine and a total of 1000 requests per user per hour, with an average of 10000 users in a peak hours and 100 in off-peak

hours with 6 user bases located in different locations around the world and 6 Data Centers also located in different locations around the world as shown in figure 5.

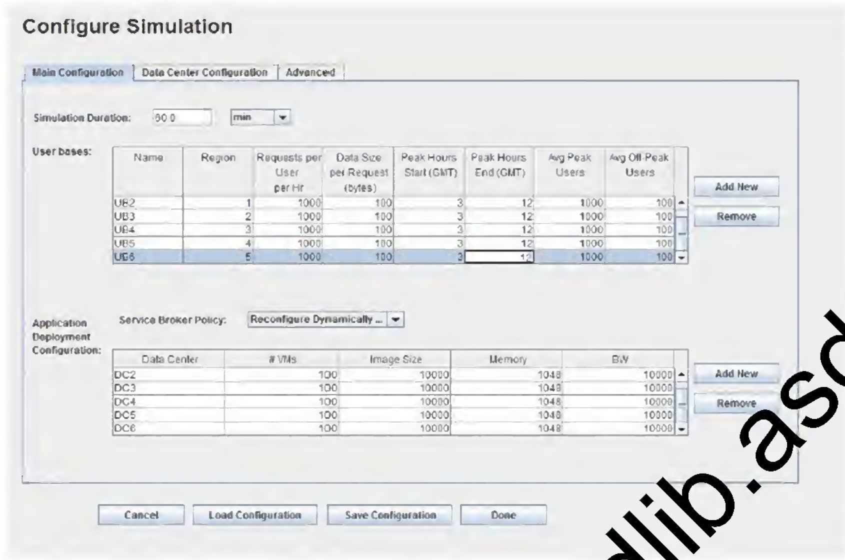

Figure 5: User base specification and service broker policy for the data center

In figure (6, 7), explanation on how to do the configuration for the simulation that is going to be tested on, Whereby the selection of the user requests that can be made per hour and whether it's a peak hour or normal hours to simulate a real world events as well as the starting and ending times for these requests. The data center broker that will govern the behavior of the data center have been identified as well ,for example in this simulation ,optimize response time which means the best data center that responded to the request to be process around the world based on the hardware specifications of the data centers like the memory , CPU speed , the number of cores inside each CPU ,the operating system, the number of virtual machines inside each data centers and the cost that for each virtual machine to use the memory and CPU cores have been defined as well .

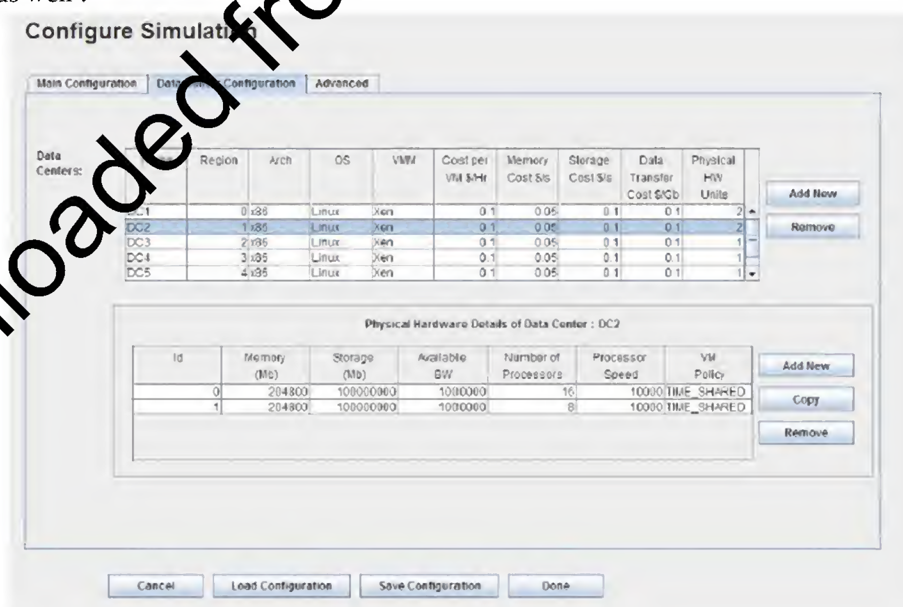

Figure 6: Data Center specifications

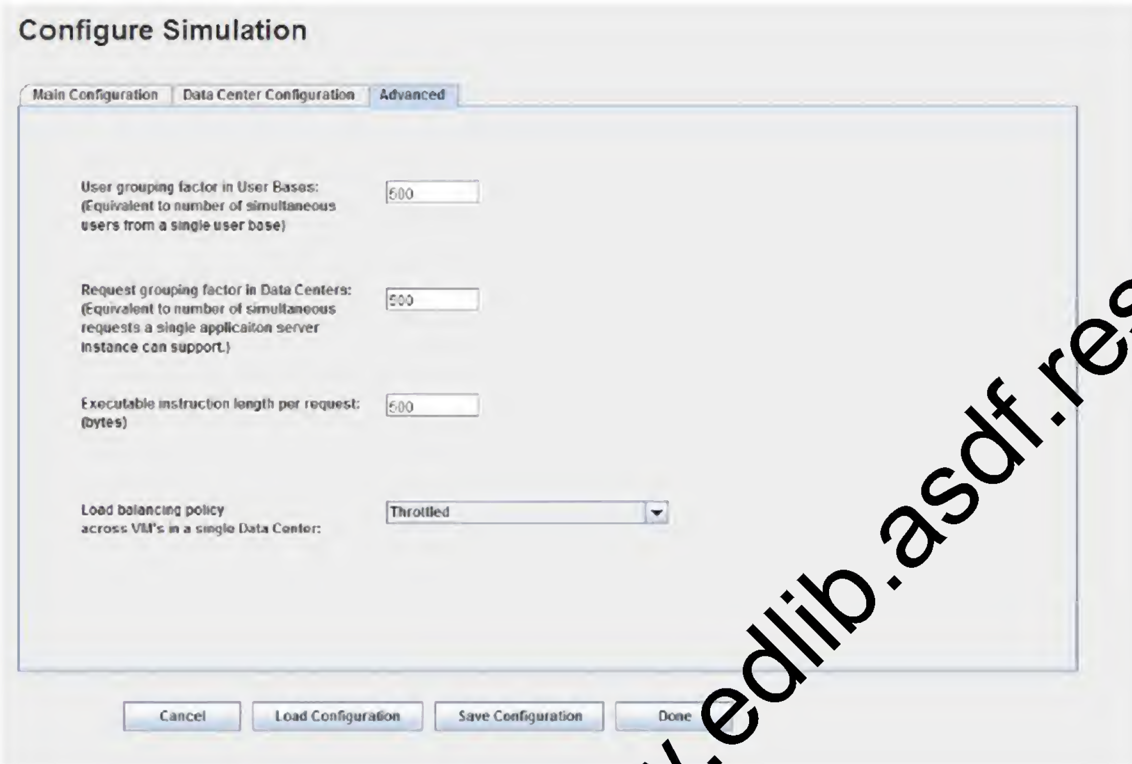

Figure 7: User requests and the load balancing algorithm

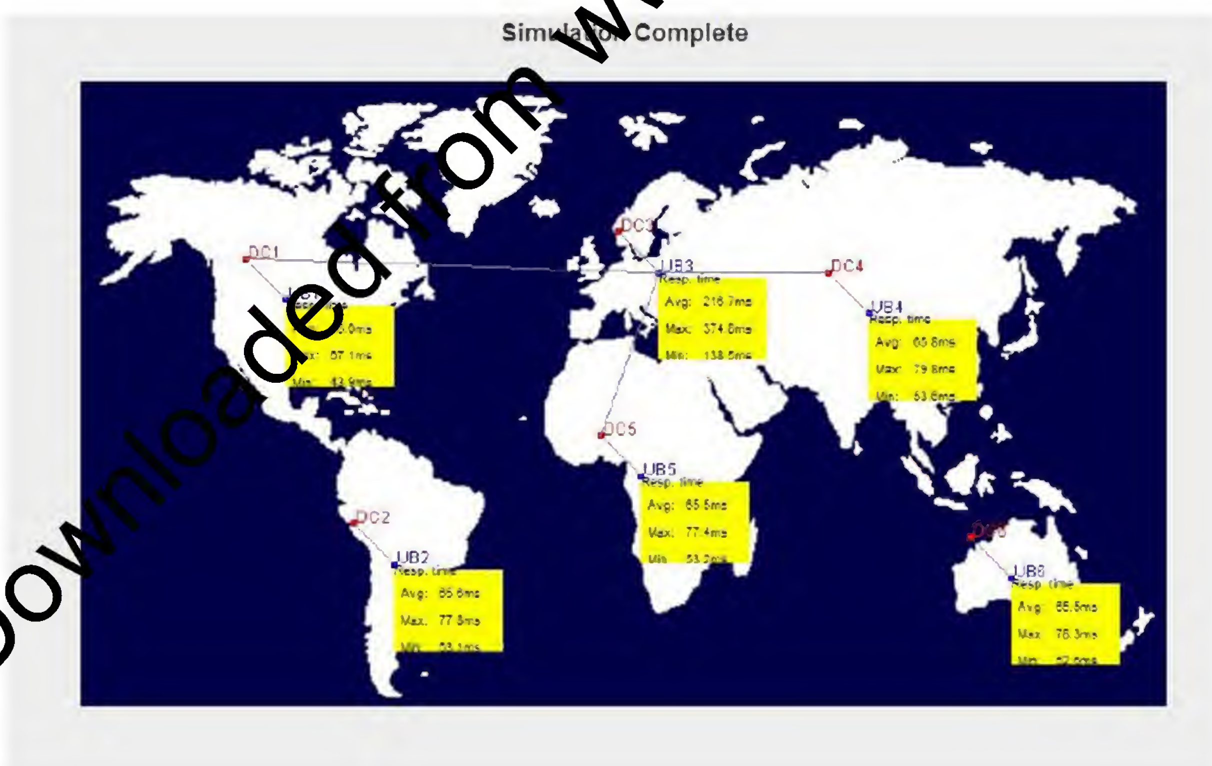

Figure 8: Data center locations and user bases requests

In figure 8, it shows the geographical location of each data center with their average, minimum and maximum execution time that can be taken in consideration when simulating with respect to each user request and whether this request is made from single user or multiple.

The results for the simulation conducted based on the specification provided for the data centers and the virtual machines shows a better results if the broker set to optimize response time in terms of data center execution time and cost of the virtual machines.

Results

Avg (ms)	Min (ms)		Max (ms)
Overall response time	89.14	43.86	374.83
Data Center processing time	26.43	3.31	162.01

Cost

Total Virtual Machine Cost ($)	Total Data Transfer Cost ($)	Grand Total ($)
50.50	5.76	56.26

Throttled with Reconfigure Dynamically

The third round was tested using the Throttled algorithm for the load balancing strategy with the Closet Data Center as the Data Center broker policy. I have conducted a simulation for 100 virtual machine and a total of 1000 requests per user per hour, with an average of 10000 users in a peak hours and 100 in off-peak hours with 6 user bases located in different locations around the world and 6 Data Centers also located in different locations around the world as shown in figure 9.

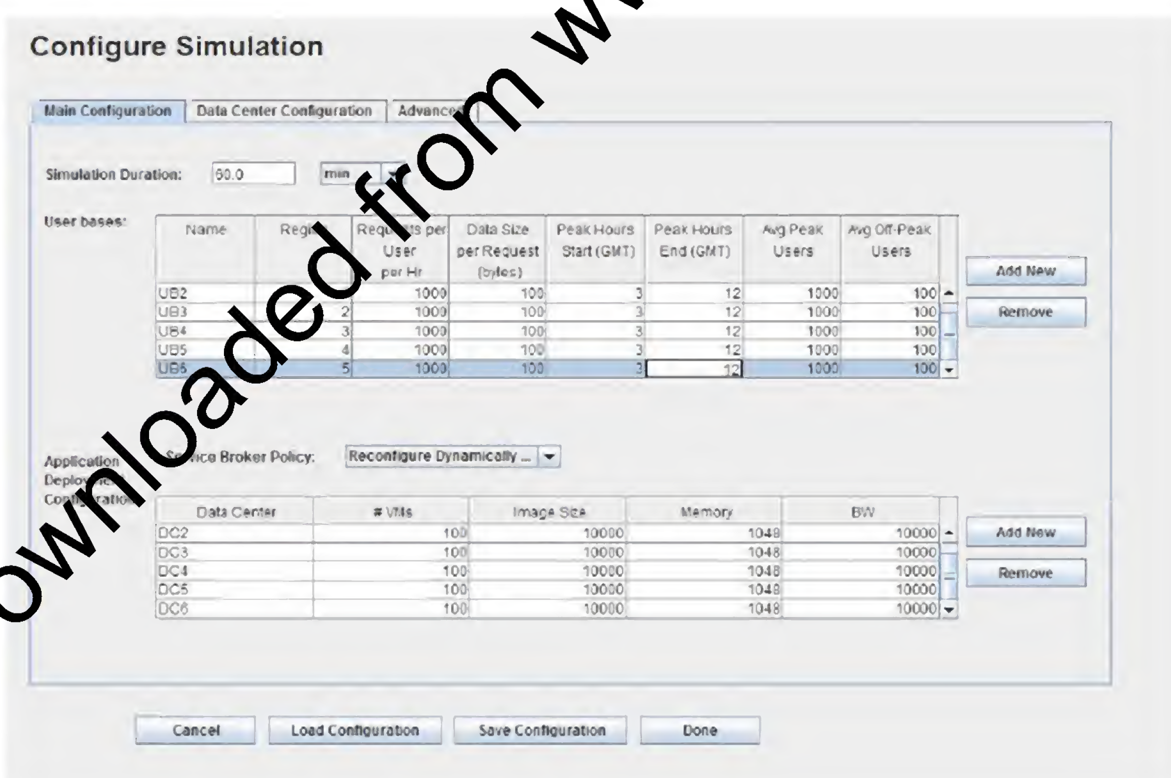

Figure 9: User base specification and service broker policy for the data center

In figure (10, 11), explanation on how to do the configuration for the simulation that is going to be tested on, Whereby the selection of the user requests that can be made per hour and whether it's a peak hour or normal hours to simulate a real world events as well as the starting and ending times for these requests. The data center broker that will govern the behavior of the data center have been identified as well ,for example in this simulation ,reconfiguration dynamically which means the data centers can serve the requests respectively but if there are other data centers available which are better than the current one, it will switch automatically to the best one accordingly. Hardware specifications of the data centers like the memory , CPU speed , the number of cores inside each CPU ,the operating system, the number of virtual machines inside each data centers and the cost that for each virtual machine to use the memory and CPU cores have been well-defined as well .

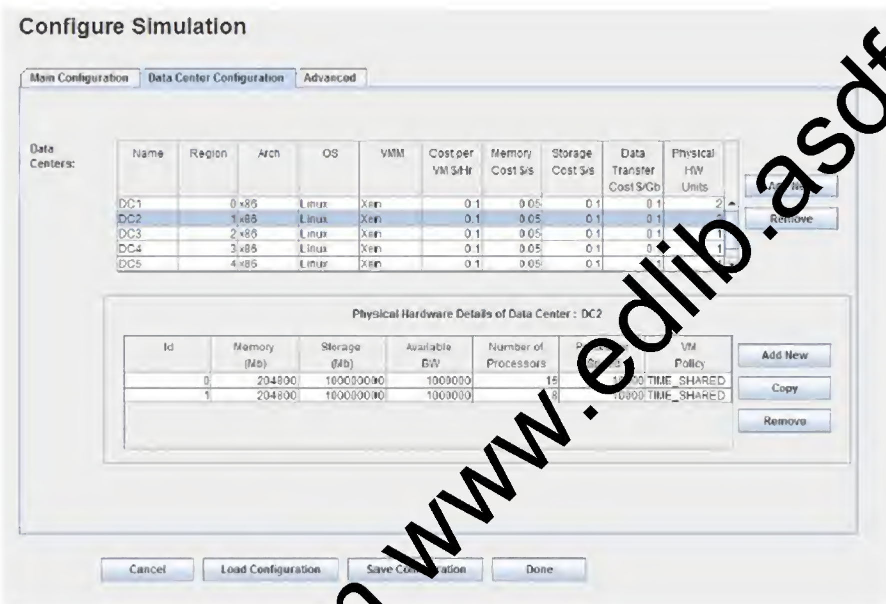

Figure 10: Data Center specifications

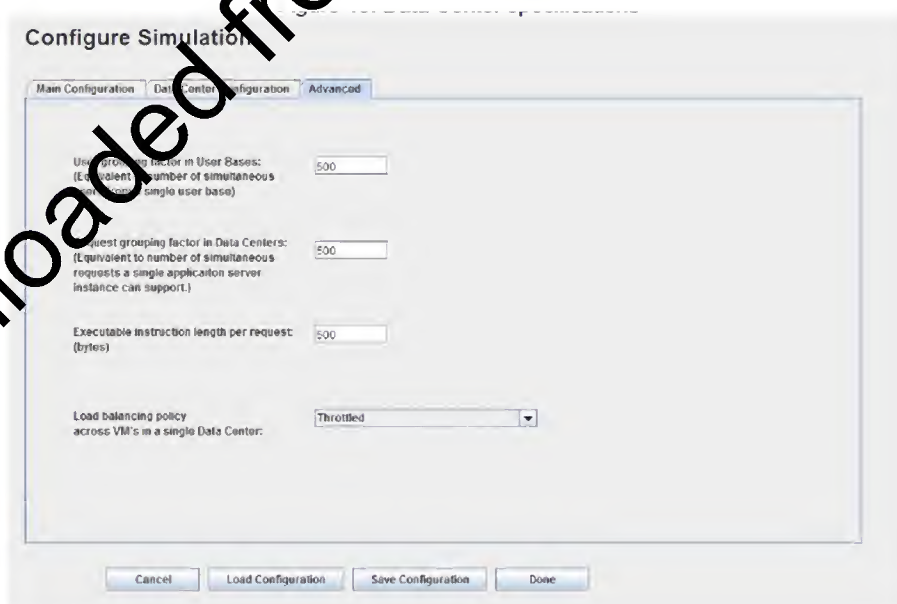

Figure 11: User requests and the load balancing algorithm

Figure 12: Data center locations and user bases requests

In figure 12, it shows the earthly location of each data center with their average, minimum and maximum execution time that can be taken in consideration when simulating with respect to each user request and whether this request is made from single user or multiple.

The results for the simulation conducted based on the specification provided for the data centers and the virtual machines shows a higher results if the broker set to reconfigure dynamically in terms of data center execution time and cost of the virtual machines than the optimize response time.

Results

Avg (ms)	Min (ms)		Max (ms)
Overall response time	139.98	44.29	227.18
Data Center processing time	90.35	3.90	169.44

Cost

Total Virtual Machine Cost ($)	Total Data Transfer Cost ($)	Grand Total ($)
53.47	5.76	59.50

Conclusion

Cloud computing is a new edge technology that is versatile, fast and developing at a fast rate. In this report, different models of implementation of cloud computing has been studied. It is obvious that the cloud concept is a way to-go method of technology implementation these days. The trend of development is high, though much has been accomplished there is still a lot to do in this field of cloud computing for the future generation. One of the biggest buzz terms in technology today is cloud computing? Companies all over the world are utilizing the cloud for their businesses, allowing users to access their technology anytime, anywhere. Essentially, organizations who are using the cloud, or cloud computing, have their files, software, information and resources available anywhere in a virtual network. A modification for throttled algorithm has been identified to increase the efficiency of its response time and data center processing execution time and cost. Implementation of this modification in the CloudSim and Cloud Analyst will conducted as a

future work. Cloud analyst has been chosen to simulate the current throttled algorithm in respect with different service broker. The results show that throttled algorithm is the chosen algorithm to modify as it has better results in terms of overall response time and data center processing.

Acknowledgement

The special thank goes to our helpful supervisor Dr. Arash Habibi Lashkari from Postgraduate school for his unrivaled supervision and guidance in our dissertation and project.

References

1. Bhadani, A., & Chaudhary, S., Performance evaluation of web servers using central load balancing policy over virtual machines on cloud. *Proceedings of the Third Annual ACM Bangalore Conference on - COMPUTE '10*, 1–4. doi:10.1145/1754288.1754304, (2010).
2. Gonçalves, G. E., Endo, P. T., Palhares, A. A., Santos, M. A., Kelner, J., & Sadok, D. (2013). On the load balancing of virtual networks in distributed clouds. *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*, 625. doi:10.1145/2480362.2480481 (2013).
3. Larosa, Y. T. Optimal QoS load balancing mechanism for virtual machines scheduling in eucalyptus cloud computing platform. *2012 2nd Baltic Congress on Future Internet Communications*, 214–221. doi:10.1109/BCFIC.2012.6217949, (2012).
4. Muhammad Shiraz, Abdullah Gani, Rashid Hafeez Khokhar, Ejaz Ahmed, An Extendable Simulation Framework for Modeling Application Processing Potentials of Smart Mobile Devices for Mobile Cloud Computing, Proceedings of Frontiers of Information Technology 2012, Pakistan, 19-21 December 2012
5. Parhizkar, B., Naser, A., Abdulhussein, A., & Joshi, H. H. (2013). A Common Factors Analysis on cloud computing models, *10*(2), IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784 523–529. March (2013).
6. Ren, X., Lin, R., & Zou, H. A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast. *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, 220–224. doi:10.1109/CCIS.2011.6045063, (2011).
7. Rajguru, A. A., & Apte, S. S. A Comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters, International journal of recent Technology and Engineering (IJRTE) , ISSN: 2277-3878, Volume-1 , Issue -3. Ausgust (2012).
8. Sethi, S., Sahu, A., & Kumar, S. (2012). Efficient load Balancing in Cloud Computing using Fuzzy Logic, *2*(7), 65–71, IOSR Journal of Engineering (IOSRJEN) ISSN 2250-3021 Volume 2, Issue 7(July 2012), PP 65-71 ,July 2012
9. Sun, P., & Chen, Y. POSTER : LBMS : Load Balancing based on Multilateral Security in Cloud, Proceedings of the 18th ACM conference on computer and communications security pages 861-864, (2011).
10. Sharma, S., Singh, S., & Sharma, M. Performance Analysis of Load Balancing Algorithms, World Academy of Science, Engineering and Technology 269–272. (2008).
11. Wang S. Yan k Liao W Wang S. Towards a Load Balancing in a Three-level Cloud Computing Network, Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on (Volume:1). (2010).
12. Wu, H. Wang C. XIE J., TeraScaler ELB-an Algorithm of Prediction-based Elastic Load Balancing Resource Management in Cloud Computing, 27th International Conference on Advanced Information Networking and Applications Workshops (2013).
13. Zohreh Sanaei, Saeid Abolfazli, Abdullah Gani, Rashid Hafeez Khokhar, Tripod of Requirements in Horizontal Heterogeneous Mobile Cloud Computing. 1st International Conference on Computing, Information Systems and Communications (CISCO'12), May 14-16, Singapore, (2012).
14. Zhang Z. and Zhang X., A Load Balancing Mechanism Based on Ant Colony and Complex Network Theory in Open Cloud Computing Federation, 2nd international conference on industrial mechatronics and automation. (2010).